PERBANDINGAN ALGORITMA JARO-WINKLER DISTANCE DAN LEVENSHTEIN DISTANCE DALAM MENDETEKSI KEMIRIPAN DOKUMEN BAHASA INDONESIA

Diajukan Untuk Menyusun Skripsi

di Jurusan Teknik Informatika Fakultas Imu Komputer UNSRI

Oleh:

NISVA SYAKBANIA

NIM: 09021181320023

Jurusan Teknik Informatika

FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA

2020

LEMBAR PENGESAHAN TUGAS AKHIR

PERBANDINGAN ALGORITMA JARO-WINKLER DISTANCE DAN LEVENSHTEIN DISTANCE DALAM MENDETEKSI KEMIRIPAN DOKUMEN BAHASA INDONESIA

Oleh :

NISVA SYAKBANIA
NIM : 09021181320023

Palembang, Agustus 2020

Pembimbing I

Novi Yusliani, M.T.
NIP. 198211082012122001

Pembimbing II,

Osvari Arsalan,M.T.
NIP. 198806282018031001

Mengetahui,
Ketua Jurusan Teknik Informatika,

Rifkie Primartha, M.T.
NIP. 197706012009121004

TANDA LULUS UJIAN SIDANG SKRIPSI

Pada hari Rabu tanggal 29 Juli 2020 telah dilaksanakan ujian sidang skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Nisva Syakbania
NIM : 09021181320023
Judul : Perbandingan Algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance* dalam Mendeteksi Kemiripan Dokumen Bahasa Indonesia

1. Pembimbing I

   Novi Yusliani, M.T.
   NIP. 198211082012122001

2. Pembimbing II

   Osvari Arsalan, M.T.
   NIP. 198806282018031001

3. Penguji I

   Samsuryadi, M.Kom., Ph.D.
   NIP. 197102041997021003

4. Penguji II

   Kanda Januar Miraswan, M.T.
   NIP. 199001092019031012

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Nisva Syakbania

NIM : 09021181320023

Program Studi : Teknik Informatika

Judul Skripsi : Perbandingan Algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance* dalam Mendeteksi Kemiripan Dokumen Bahasa Indonesia

Hasil Pengecekan Software *iThenticate/Turnitin* : 2 %

Menyatakan bahwa Laporan Proyek saya merupakan hasil karya saya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan proyek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Palembang, Agustus 2020

Nisva Syakbania
NIM. 09021181320023

MOTTO DAN PERSEMBAHAN

- “Beginning always the hardest.”
- “Today, you are you that is truer than true. There is no one alive that is youer than you.” (dr. Seuss)
- “Sometimes the questions are complicated and the answers are simple.” (dr. Seuss)
- “There is no greater weapon than a prepared mind.” (Zhuge Liang)
- “This too shall pass.”

Kupersembahkan karya tulis ini kepada :

- Myself,
- My dear parents,
- My ever-so-great sister,
- My beloved friends,
- Almamater.

PERBANDINGAN ALGORITMA JARO-WINKLER DISTANCE DAN LEVENSHTEIN DISTANCE DALAM MENDETEKSI KEMIRIPAN DOKUMEN BAHASA INDONESIA

By:
Nisva Syakbania
09021181320023

ABSTRACT

Document similarity detection is used to calculate the similarity between two or more documents based on semantic similarity or lexical similarity. This research proposed to detect similarity based on lexical similarity using a string matching techniques on each documents. Jaro-Winkler and Levenshtein Distance are algorithms usually used in string matching techniques. Jaro-Winkler Distance includes a step of calculating the length of strings in the document, counting common characters, and transposition. Levenshtein Distance is an algorithm which is used to calculate the minimum distance that needed to transform one string into the other. Testing was done with a total 19 authentic document and 6 comparative, the result of this research shows that the average error value of Levenshtein Distance is 7,86% while Jaro-Winkler Distance with average error value of 24,45%. As for computing time, four out of five testing configuration shows that Jaro-Winkler Distance have a faster computing time than Levenshtein Distance.

Keywords : Document Similarity, Jaro-Winkler Distance, Levenshtein Distance

PERBANDINGAN ALGORITMA JARO-WINKLER DISTANCE DAN LEVENSHTEIN DISTANCE DALAM MENDETEKSI KEMIRIPAN DOKUMEN BAHASA INDONESIA

Oleh:
Nisva Syakbania
09021181320023

ABSTRAK

Deteksi kemiripan dokumen digunakan untuk mengetahui kemiripan antara dua buah dokumen yang biasanya diukur berdasarkan kesamaan semantik atau kesamaan leksikal. Pada penelitian ini dilakukan deteksi kemiripan secara leksikal dengan cara melakukan pencocokan string pada setiap dokumen. Algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance* merupakan algoritma yang sering digunakan dalam teknik pencocokan pola. Algoritma *Jaro-Winkler Distance* menghitung panjang kata dalam dokumen, kata yang sama, dan jumlah transposisi. Sedangkan algoritma *Levenshtein Distance* menghitung jarak yang dibutuhkan untuk mengubah satu kata menjadi kata lain. Pengujian dilakukan dengan menggunakan 19 dokumen asli dan 6 dokumen pembanding menghasilkan rata-rata nilai error sebesar 7,86% untuk algoritma *Levenshtein Distance* dan 24,45% untuk algoritma *Jaro-Winkler Distance.* Untuk waktu proses, empat dari lima skenario pengujian menunjukkan *Jaro-Winkler Distance* memiliki waktu komputasi yang lebih cepat daripada *Levenshtein Distance.*

Kata Kunci: Kemiripan Dokumen, *Jaro-Winkler Distance, Levenshtein Distance*

KATA PENGANTAR

Segala puji dan syukur bagi Allah SWT atas berkah dan rahmat-Nya yang telah diberikan kepada Penulis sehingga dapat menyelesaikan Tugas Akhir ini dengan baik. Tugas akhir yang berjudul **"Perbandingan Algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance* dalam Mendeteksi Kemiripan Dokumen Bahasa Indonesia"** ini disusun untuk memenuhi salah satu syarat guna menyelesaikan pendidikan program Strata-1 pada Fakultas Ilmu Komputer Program Studi Teknik Informatika di Universitas Sriwijaya.

Pada kesempatan ini Penulis ingin menyampaikan ucapan terima kasih kepada pihak-pihak yang telah memberikan dukungan, bimbingan, dan motivasi selama proses penelitian ini dilaksanakan. Secara khusus Penulis ingin menyampaikan terima kasih kepada:

1. Allah SWT yang selalu menjawab doa hamba-Nya.
2. Nisva Syakbania, terima kasih karena tidak menyerah dan selalu percaya pada diri sendiri bahwa ia bisa menyelesaikan tugas akhirnya.
3. Kedua orang tua tercinta, terima kasih untuk selalu menyebut namaku dalam setiap doa. Keyakinan dan dukungan moril maupun materil yang tiada henti telah membantu Penulis dalam menyelesaikan Tugas Akhir ini.
4. Saudaraku, Mega Puspita, atas dukungan dan kata-kata bijak yang telah membangkitkan motivasi Penulis.
5. Bapak Jaidan Jauhari, M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.

6. Bapak Rifkie Primartha, M.T. selaku Ketua Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Ibu Novi Yusliani, M.T. selaku dosen pembimbing I dan Bapak Osvari Arsalan, M.T. selaku dosen pembimbing II yang telah membimbing, mengarahkan, dan memberikan motivasi untuk Penulis dalam proses pengerjaan Tugas Akhir.
8. Bapak Samsuryadi, M.Kom., Ph.D. dan Bapak Kanda Januar, M.T. selaku penguji Tugas Akhir yang telah memberi nasihat dan saran yang membangun.
9. Bapak Julian Supardi dan Bapak Danny Matthew Saputra, M.Sc. selaku pembimbing akademik.
10. Seluruh dosen Program Studi Teknik Informatika yang telah memberikan bekal ilmu selama masa perkuliahan.
11. Sahabat seperjuangan Teknik Informatika Angkatan 2013.
12. Seluruh staff administrasi dan pengawai yang telah membantu dalam urusan administrasi

Penulis menyadari masih terdapat banyak kekurangan dalam penyusunan Tugas Akhir disebabkan oleh keterbatasan pengetahuan dan pengalaman. Oleh karena itu kritik dan saran yang membangun sangat diharapkan untuk menyempurnakan Tugas Akhir ini. Semoga Tugas Akhir ini dapat bermanfaat bagi kita semua.

Palembang, Agustus 2020

Nisva Syakbania

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN

Halaman

BAB I

PENDAHULUAN

1.1 Pendahuluan

Pada bab ini membahas secara sistematis mengenai penelitian tugas akhir yang meliputi latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, dan batasan masalah yang menjelaskan batasan perangkat lunak yang akan dibuat.

1.2 Latar Belakang Masalah

Perkembangan pesat di dunia teknologi dan informasi memudahkan setiap kalangan untuk mengakses berbagai informasi. Seiring dengan kemudahan akses informasi ini terdapat juga beberapa dampak negatif, salah satunya praktik pelanggaran etika akademis yang dilakukan oleh kalangan akademisi. Para akademisi ini dapat dengan mudah mengambil data dan informasi yang tersebar di internet lalu menggunakannya tanpa mencantumkan sumber. Hal ini dibuktikan dengan banyak ditemukannya kasus plagiarisme oleh mahasiswa dan tenaga pengajar di universitas.

Kemiripan teks merupakan salah satu bentuk plagiarisme dimana aspek yang ditiru berupa kata demi kata, kalimat, paragraf atau bahkan keseluruhan karya orang lain (Sastroasmoro, 2007). Pendeteksian kemiripan ini dapat dilakukan dengan melakukan perbandingan antara teks asli dengan teks yang dicurigai.

Apabila tingkat kemiripan teks mencapai persentase tertentu (tergantung kebijakan tiap instansi), maka karya tulis tersebut dapat dinyatakan sebagai plagiarisme.

Salah satu langkah yang dapat dilakukan untuk mendeteksi kemiripan teks adalah dengan deteksi secara manual dan otomatis. Deteksi kemiripan dokumen secara otomatis sendiri sudah banyak dikembangkan dalam beberapa tahun ini. Deteksi kemiripan teks dengan membandingkan keseluruhan isi dokumen kata per kata dapat dilakukan dengan pendekatan pencocokan string, salah satunya *approximate string matching.* Metode ini merupakan teknik dalam pencocokan pola pada string berdasarkan kemiripan tekstual atau penulisan meliputi jumlah karakter dan susunan karakter dalam dokumen (Rochmawati & Kusumaningrum, 2016). Contoh algoritma yang sering digunakan dalam pendekatan ini adalah algoritma *brute force, boyer-moore, jaro-winkler, levenshtein distance,* dan *hamming distance*.

Pendeteksian kemiripan dokumen bahasa Indonesia pada penelitian ini akan menggunakan algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance.* Algoritma *Jaro-Winkler Distance* adalah algoritma untuk menghitung kesamaan antara dua string dan merupakan salah satu algoritma yang paling efektif dalam menghitung tingkat kemiripan. Dalam penelitian perbandingan metode *Approximate String Matching* dalam mengidentifikasi kesalahan penulisan (Rochmawati & Kusumaningrum, 2016), algoritma *Jaro-Winkler Distance* merupakan algoritma yang memiliki nilai terbaik dalam melakukan perbaikan ejaan pada bahasa Indonesia dengan nilai MAP *(Mean Average Precision)* sebesar 0,87 dibandingkan algoritma lainnya. Dalam penelitian yang dilakukan oleh (Tinaliah &

Elizabeth, 2018) yang berjudul "*Perbandingan Hasil Deteksi Plagiarisme Dokumen dengan Metode Jaro-Winkler Distance dan Metode Latent Semantic Analysis*", metode Jaro-Winkler memberikan hasil plagiat 100% pada data yang sama persis dibandingkan dengan metode *LSA* yang menghasilkan nilai kemiripan 97,14%.

Levenshtein Distance merupakan algoritma pengukuran string untuk menghitung perbedaan antara dua kata dimana semakin kecil nilai yang dimiliki antara dua kata tersebut maka semakin tinggi tingkat kemiripan. Algorima *Levenshtein Distance* sudah digunakan di berbagai bidang seperti perbaikan ejaan, mesin pencari, dan juga deteksi plagiarisme. Berdasarkan hasil penelitian terdahulu, didapatkan hasil akurasi sebesar 86% dalam pengoreksian kesalahan masukan pada sistem pencarian (Yulianto, Arifudin, & Alamsyah, 2018).

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah dikemukakan sebelumnya, maka penelitian tugas akhir yang dilakukan adalah mendeteksi kemiripan dokumen Bahasa Indonesia menggunakan algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance.* Untuk mendukung rumusan masalah tersebut, maka penelitian ini akan dibagi dalam *research questions,* yaitu:

1. Berapa lama waktu proses deteksi kemiripan dokumen bahasa Indonesia dengan algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance.*

2. Berapa nilai kemiripan yang dihasilkan algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance* dalam mendeteksi kemiripan dokumen bahasa Indonesia.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Untuk mengetahui waktu proses algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance* dalam mendeteksi kemiripan dokumen bahasa Indonesia.
2. Untuk mengetahui perbedaan persentase kimiripan dokumen yang dihasilkan dari algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance.*

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah:

1. Hasil penelitian ini dapat mendeteksi kemiripan antara dua dokumen yang kemudian dapat digunakan sebagai dugaan awal dalam deteksi plagiat.
2. Hasil penelitian ini dapat digunakan sebagai referensi untuk pemilihan algoritma yang lebih baik untuk mendeteksi kemiripan dokumen.

1.6 Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Dataset yang digunakan merupakan dokumen dalam bahasa Indonesia.
2. Format dokumen masukan adalah *.txt dan *.pdf.

3. Penelitian ini hanya mengukur persentase kemiripan antar dua buah dokumen.

1.7 Sistematika Penulisan

Sistematika penulisan skripsi ini adalah sebagai berikut:

BAB I. PENDAHULUAN

Pada bab ini diuraikan mengenai latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah, metodologi penelitian, dan sistematika penulisan.

BAB II. KAJIAN LITERATUR

Pada bab ini akan membahas landasan teori yang digunakan dalam penelitian ini, seperti definisi kemiripan teks, analisis *preprocessing*, analisis algoritma *Jaro-Winkler Distance* dan algoritma *Levenshtein Distance.* Selain itu akan dibahas mengenai penelitian-penelitian terkait yang penelitian ini.

BAB III. METODOLOGI PENELITIAN

Pada bab ini akan dibahas mengenai tahapan yang akan dilaksanakan pada penelitian ini. Masing-masing rencana tahapan penelitian dideskripsikan dengan rinci berdasarkan pada kerangka kerja. Di akhir bab akan dijabarkan perancangan manajemen proyek perangkat lunak untuk pelaksanaan penelitian ini.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Pada bab ini akan dibahas mengenai analisis, perancangan dan lingkungan implementasi perangkat lunak deteksi kemiripan dokumen bahasa Indonesia, implementasi algoritma *Jaro-Winkler Distance*, implementasi algoritma *Levenshtein Distance*, hasil eksekusi, dan hasil pengujian.

BAB V. HASIL DAN ANALISIS PENELITIAN

Pada bab ini pengujian dilakukan berdasarkan skenario-skenario pengujian yang telah dirancang sebelumnya. Kemudian dilakukan analisis sebagai basis dari kesimpulan yang diambil dalam penelitian ini.

BAB VI. KESIMPULAN DAN SARAN

Pada bab ini berisi kesimpulan dari semua uraian-uraian pada bab-bab sebelumnya dan juga berisi saran-saran yang diharapkan berguna dalam penerapan perangkat lunak deteksi kemiripan dokumen bahasa Indonesia di penelitian selanjutnya.

1.8 Kesimpulan

Penelitian ini berfokus pada deteksi kemiripan dokumen berdasarkan kemiripan penulisan tekstual pada dokumen bahasa Indonesia, serta untuk membandingkan dua algoritma untuk melihat algoritma mana yang lebih baik dalam mendeteksi kemiripan dokumen. Algoritma yang digunakan adalah

algoritma *Jaro-Winkler Distance* dan *Levenshtein Distance.* Dokumen yang digunakan adalah dokumen bahasa Indonesia dengan esktensi *.pdf* dan *.txt.* Hasil akhir yang diharapkan dari penelitian ini adalah mengetahui perbedaan persentase kemiripan dan waktu proses yang dihasilkan oleh kedua algoritma tersebut.

DAFTAR PUSTAKA

Adiwidya, B. M. D. (2009). Algoritma Levenshtein Dalam Pendekatan Approximate.

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., & Williams, H. E. (2007). Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing*, *6*(4).

Agusta, L. (2009). Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia. *Konferensi Nasional Sistem Dan Informatika 2009*, (KNS&I09-036), 196–201.

Friendly, F. (2019). Jaro-Winkler Distance Improvement for Approximate String Search Using Indexing Data for Multiuser Application. *Journal of Physics: Conference Series*, *1361*(1).

Gomma, W. H., & Fahmy, A. A. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, *68*(13), 13–18.

Kurniawan, W., Bijaksana, M. A., & Wahyudi, B. A. (2018). Analisis Pencocokan Nama dengan Nama Arab Terjemahan Bahasa Indonesia Menggunakan metode Levenshtein Distance Pendahuluan Studi Terkait Hadits, *5*(3), 7472–7493.

Kurniawati, A. (2010). Implementasi Algoritma Jaro-Winkler Distance untuk

Membandingkan Kesamaan Dokumen Berbahasa Indonesia. *Proceeding, Seminar Ilmiah Nasional Komputer Dan Sistem Intelijen KOMMIT 2008, Depok, Indonesia*.

Rochmawati, Y., & Kusumaningrum, R. (2016). Studi Perbandingan Algoritma Pencarian String dalam Metode Approximate String Matching untuk Identifikasi Kesalahan Pengetikan Teks. *Jurnal Buana Informatika, 7*(2), 125–134.

Sastroasmoro, S. (2007). Beberapa Catatan tentang Plagiarisme *. *Maj Kedokt Indon, Volum: 57*, 239–244.

Su, Z., Ahn, B. R., Eom, K. Y., Kang, M. K., Kim, J. P., & Kim, M. K. (2008). Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm. *3rd International Conference on Innovative Computing Information and Control, ICICIC'08*, 0–3.

Tinaliah, T., & Elizabeth, T. (2018). Perbandingan Hasil Deteksi Plagiarisme Dokumen dengan Metode Jaro-Winkler Distance dan Metode Latent Semantic Analysis. *Jurnal Teknologi Dan Sistem Komputer*, *6*(1), 7.

Yulianto, M. M., Arifudin, R., & Alamsyah, A. (2018). Autocomplete and Spell Checking Levenshtein Distance Algorithm To Getting Text Suggest Error Data Searching In Library. *Scientific Journal of Informatics*, *5*(1), 75.